# Deep Reinforcement Learning (Sp25)

**Instructor: Dr. Mohammad Hossein Rohban**

**Midterm Exam | 1404/01/21 | 120 Minutes**

## 1 Introduction to RL

### 1.1 Question

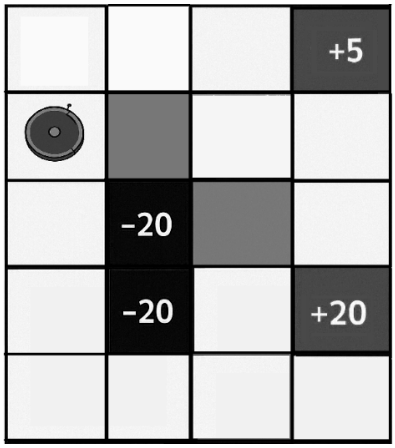
Consider the gridworld environment shown below. In all states, the agent has available actions ↑, ↓, ←, →. Performing an action that would transition to an invalid state (outside the grid or into a wall) results in the agent remaining in its original state. The vacuum cleaner robot starts in the cell (2,1) and has two exits:

1. A close one to collect some small trash with a +5 reward.

2. A distant one with a +20 reward Stepping into red cells gives a –20 penalty (a hazardous wet area that can damage the vacuum). Gray cells are walls.

The agent behaves differently under parameter settings, depending on:

- $\gamma$: the discount factor (how much future reward is valued)
- `noise`: 0 = deterministic, 0.5 = stochastic




1. Explain Markov property and define it for the gridworld example.
2. Write different components of a MDP for the gridworld environment when noise = 0.
3. Match each behavior (a–c) with one of the parameter sets (1–3) , explain your answer.

**Agent Behaviors:**

(a) Prefers the **close exit** $(+5)$
(b) Prefers the **distant exit** $(+20)$, **risking** the wet area $(-20)$
(c) Prefers the **distant exit** $(+20)$, **avoiding** the wet area $(-20)$

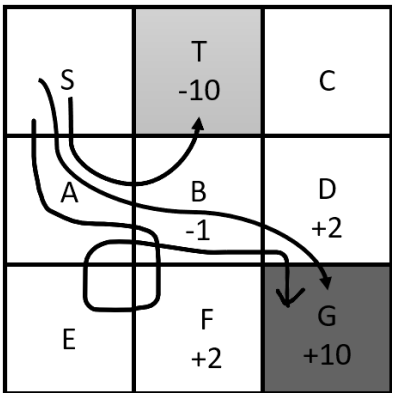
**Parameter Settings:**

1. $\gamma = 0.1$, noise $= 0.5$
2. $\gamma = 0.99$, noise $= 0.5$
3. $\gamma = 0.99$, noise $= 0$

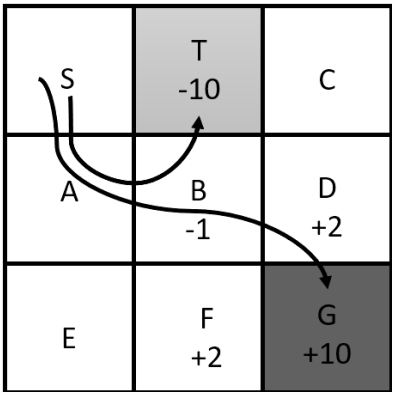
## 2 Value-Based Methods

### 2.1 Question

**a) Compute the value of state B using both first visit and every visit MC (Consider $\gamma$ = 1).**




**b) Compute the value of states S, B, A, and D using one-step TD-learning (Consider $\gamma$ = 1 and $\alpha$ = 0.5).**




### 2.2 Question

**An agent implements Q-learning with experience replay in a non-stationary environment. If the replay buffer contains 10,000 transitions from older versions of the environment, explain the effect this would have on the convergence properties compared to standard Q-learning without replay.**

### 2.3 Question

**a) Explain the (Bias/Variance) trade-off in MC and TD-learning methods.**
**b) Explain the structure of the replay buffer and mention two reasons why we use it.**

# 3 Policy-Based Methods

## 3.1 Question

Let $\pi_\theta(a \mid s)$ be a differentiable stochastic policy with parameters $\theta$, and consider the objective function defined as the expected return over trajectories:

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[R(\tau)\right], \quad \text{where} \quad R(\tau) = \sum_{t=0}^{T} r(s_t, a_t)$$

Provide a complete and well-explained derivation of the gradient $\nabla_\theta J(\theta)$ in terms of the policy $\pi_\theta$, starting from the definition of the expected return. Include all necessary steps and ensure the explanation is clear and correct.

**Additionally**, clearly state all mathematical assumptions required for the derivation to be valid.

## 3.2 Question

What are the two major sources of variance in the REINFORCE estimator? Briefly explain how introducing a baseline helps in reducing this variance without introducing bias.

# 4 Advanced Methods

## 4.1 Question

In Soft Actor-Critic, the policy improvement step involves maximizing the expected entropy-regularized reward:

$$J_\pi(\theta) = \mathbb{E}\left[Q(s,a) - \alpha \log \pi(a|s)\right] \tag{1}$$

where $\alpha$ is the entropy temperature coefficient.

**(a)** Explain why the entropy term $-\alpha \log \pi(a|s)$ is included in the objective function and how it affects exploration.

**(b)** Suppose for a given state $s$, the Q-values and policy probabilities for two actions are given as:

| Action $a$ | $Q(s,a)$ | $\pi(a\|s)$ |
|---|---|---|
| $a_1$ | 2.0 | 0.7 |
| $a_2$ | 5.5 | 0.3 |

If $\alpha = 0.1$, compute the policy improvement objective for this state.

**(c)** In the evaluation phase of SAC, where actions are selected greedily, which action should be chosen in this state?

### 4.2 Question

**What are the implications of modifying PPO's clipping range from the standard $[1-\epsilon, 1+\epsilon]$ to either $[1-\epsilon, 1+2\epsilon]$ or $[0, 1+\epsilon]$? Discuss how these modifications affect learning stability, exploration.**

## 5 Model-Based Methods

### 5.1 Question

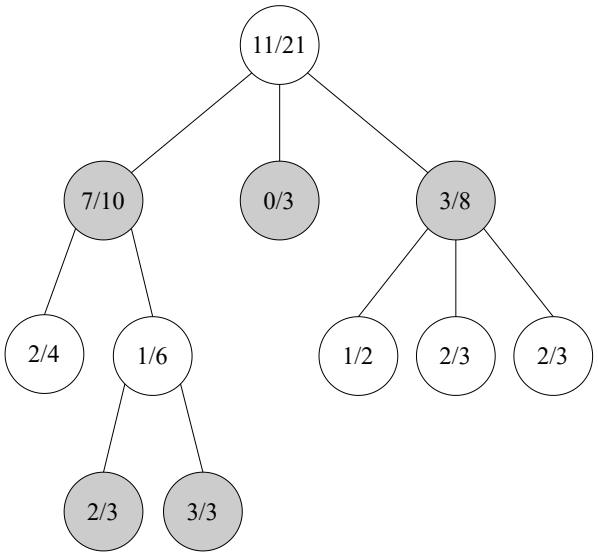
**Answer the following questions concisely:**

1. How does the Upper Confidence Bound (UCB) formula help balance the exploration-exploitation trade-off in Monte Carlo Tree Search (MCTS)?
2. How can the Dyna-Q algorithm be adapted for non-deterministic (stochastic) environments to improve learning and decision-making?
3. How does adding a baseline, help Q-learning and its variants? Provide 2 reasons.

### 5.2 Question

In the provided graph, two players (White and Black) compete, and the game ends in either victory or defeat (draws are not possible). Using Monte Carlo Tree Search (MCTS), perform the following tasks:

1. Determine the next node to explore based on the current state of the tree, applying the Upper Confidence Bound (UCB1) formula.
2. Assume the simulation results in a defeat for the White player. Update the nodes by propagating this result back to the root, modifying visit counts and win probabilities accordingly.
3. After backpropagation, identify the next node for exploration and explain how the updated statistics influenced your decision.



# 6 Multi-Armed Bandits

## 6.1 Question

**Give two methods for dealing with a non-stationary (the reward distributions of arms change over time) multi-armed bandit environment.**

## 6.2 Question

**Suppose an online seller wants to recommend m of their products to users. In each recommendation, they present an offer to a user, and their goal is to maximize the number of users who open the offer. Assume each user is represented by a feature vector X.**

(a) How would you solve this problem using a bandit approach? Explain the details of your solution.
(b) Suppose we know that user responses depend on special occasions (such as holidays). What modification should be made to the above algorithm to account for this challenge?
(c) The response distribution of a user may depend on the history of offers presented to them. Given this, why does regret become even more important in this problem?